Temple University

Philadelphia, Pennsylvania

Center for Statistical Distributions

1. This document has been approved for public release and sale; its distribution is unlimited.

57

Twenty-five years of progress in information theory

S. Kotz[1] and J. Wolfowitz[2]

August 1968

Twenty-five years of progress in information theory

S. Kotz[1] and J. Wolfowitz[2]

0. Introduction. Contents of this report.

We congratulate the University of California on its centenary and are pleased to contribute this report in its honor. Our title is actually a slight misnomer, since the work we shall describe begins with the 1948 paper of Shannon [141].

Information theory covers a multitude of subjects (the cynic might say sins) and we would like here briefly to indicate what this report will and will not cover. It will concern itself entirely with what is often called probabilistic coding theory. Algebraic coding theory, which could properly be considered a branch of information theory, will not be included because it is largely outside the competence of the authors. Although algebraic coding theory and probabilistic coding theory are parallel and complementary in one sense, their spirits and methods are very different. There are other mathematical disciplines which are often incorrectly lumped under information theory, principally because they use entropy function as a tool. It would be as incorrect to classify them under information theory as it would be to call any theory integration theory simply because it involves the use of the integral as a tool. Thus we shall not discuss the problems in ergodic theory which have been solved by using entropy as an invariant, nor problems of packing in function spaces, nor the entropy of stochastic processes, not the various systems of

axiomatizing entropy.

Appended to this paper is a bibliography which is reasonably complete, though not exhaustive. It is obviously impossible for us to discuss every one of these papers, particularly as the editors of this volume, of necessity, have subjected us to precise space limitations. The choice open to us was therefore either to write an introductory exposition of information theory or a very technical paper for specialists. The first of these choices seemed to us not to be in keeping with the spirit of this volume, and the second would result in a paper which could be read only by a small group who might have little need for reading it. We have therefore decided to compromise between the two choices. We shall discuss a number of basic, typical, and important subjects, which will enable the non-specialist reader to get some of the flavor and some understanding of the theory, without at the same time completely boring the specialist reader. We can only hope that this compromise will not cause us to fail on both counts.

In order to avoid invidious comparisons and for other reasons, we have decided to omit actual citation of references in the text. There are only two exceptions to this. One, a very minor one, is where we cite two papers with seemingly contradictory results, because we wish to warn the reader that they deal with different versions of a problem discussed below. The other, the major exception, is to refer freely to the name and papers of C.E. Shannon, whose truly brilliant work founded the theory and produced many of its important results.

Footnotes

1) Work supported by the Air Force Office of Scientific Research under Grant AF-AFOSR-68-1411 to Temple University.

2) Work supported by the Air Force Office of Scientific Research under Grant AF-AFOSR 396-63 to Cornell University.

1. Discrete memoryless channels.

Let $A^* = \{1,\ldots,a\}$ and $B^* = \{1,\ldots,b\}$ be, respectively, the input and output alphabets. The alphabet that we use in everyday life consists of 26 Latin letters, 10 numerical symbols, various punctuation marks, and a space between words, which is itself a punctuation mark. The alphabets A^* and B^* are essentially no different and no less general. To avoid the trivial, we assume that both a and b are greater than 1.

Any sequence of n letters, or elements, from A^* (respectively, from B^*) is called a transmitted or sent n-sequence (respectively, a received n-sequence). In any one discussion, n will be fixed. The sender transmits n-sequences over a channel. When he sends such a sequence, say u_o, the receiver receives a chance received n-sequence; that is, the sequence received depends upon chance. Call the chance received n-sequence $v(u_o)$. Its distribution depends on u_o and the channel. In fact, for mathematical purposes the channel is simply the function

$$P\{v(u_o) = v_o\}, \tag{1.1}$$

that is, the probability that, when the n-sequence u_o is sent, the chance received sequence should be v_o; this function is defined for any transmitted n-sequence u_o and any received n-sequence v_o. When necessary to avoid confusion, dependence on n should be indicated. Usually the function (1.1) is defined for every n.

One of the simplest and most important of all channels is the discrete memoryless channel (dmc). It is described by means of a channel probability function (cpf) $w(j|i)$, defined for every $i \in A^*$ and every $j \in B^*$. This can be any function for which always $w(j|i) \geq 0$ and

$$\sum_{j\in B^*} w(j|i) = 1, \qquad i \in A^*.$$

Different functions w define different dmc's. Let

$$u_o = (a_1, a_2, \ldots, a_n),$$

$$v_o = (b_1, b_2, \ldots, b_n).$$

Then

$$P\{v(u_o) = v_o\} = \prod_{k=1}^{n} w(b_k|a_k).$$

We see that $w(j|i)$ can be regarded as the probability that, when the letter i is sent, the letter j is received. In that case, the individual letters received are independently distributed.

We now define the notion of codes which is basic in information theory. A code (n,N,λ), where n is the length of each word, N is the length of the code, and λ is the maximum probability of error, is a system

$$\{(u_1,A_1),\ldots,(u_N,A_N)\}, \tag{1.2}$$

where $u_1,\ldots,u_N$ are transmitted n-sequences, $A_1,\ldots,A_N$ are disjoint sets of received n-sequences, and

$$P\{v(u_i) \in A_i\} \geq 1-\lambda, \qquad i = 1,\ldots,N. \tag{1.3}$$

A code is used as follows: When the sender wishes to send the ith message, he sends u_i. When the message received lies in A_j, the receiver concludes that the jth message was sent. If the message received does not lie in any A_j, he may draw any conclusion he wishes about the message that has been sent. The probability

that any message sent will be correctly understood by the receiver is at least $1 - \lambda$.

One general problem is this: For various channels of interest, given n and λ, $0 < \lambda < 1$, how big can N be? Most of the known results are asymptotic in n.

The closely related problem of constructing the codes whose existence is guaranteed by the theorems that will be cited below is as yet only partially solved.

Any vector with nonnegative components that add to 1 may be called a probability distribution. A probability distribution on A^* (respectively on B^*) will have a (respectively b) components.

The entropy of a probability distribution π,

$$\pi = (\pi_1, \ldots, \pi_c),$$

is defined to be

$$H(\pi) = - \sum_{i=1}^{c} \pi_i \log_2 \pi_i. \tag{1.4}$$

Logarithms to the base 2 are used for historical reasons only, and any other base would do as well. If $\pi_i = 0$, the ith term of the right-hand member of (1.4) is defined to be 0. This last convention always applies. The entropy function has many important combinatorial properties which are essential in the statement and proofs of most coding theorems.

Let $N(n,\lambda)$ be the length of the longest code (i.e., of maximum length) of word length n and maximum probability of error λ. Obviously $N(n,\lambda)$ is a monotonically non-decreasing function

of λ. Yet the following remarkable theorem holds:

$$\lim_{n\to\infty} \frac{1}{n} \log_2 N(n,\lambda) = C, \tag{1.5}$$

where C is a constant, independent of λ, given by

$$\max_{\pi} [H(\pi') - \sum_i \pi_i H(w(\cdot|i))], \tag{1.6}$$

where

$$\pi' = W \pi, \tag{1.7}$$

W is the matrix with $w(j|i)$ the element in the jth row and i^{th} column, and π and π' are probability distributions (column vectors) on A^* and B^*, respectively. The number C is called the capacity of the channel. One can say even more! There exists a positive function $K(\lambda)$ of λ such that, for any n, there exists a code such that

$$N > \exp_2 \{nC - \sqrt{n}\ K(\lambda)\} \tag{1.8}$$

and there does not exist a code such that

$$N > \exp_2 \{nC + \sqrt{n}\ K(\lambda)\}. \tag{1.9}$$

(1.8) is called a coding theorem and (1.9) is called its strong converse. The weaker result, that always

$$N(n,\lambda) < \exp_2 \{\frac{nC + 1}{1-\lambda}\} \tag{1.10}$$

is called a weak converse.

A channel other than the one has a different function (1.1), (not given by the product of the values of $w(\cdot|\cdot)$) and may have different alphabets. Whenever, for such a channel, (1.5) is satisfied, we shall call C the capacity of the channel. Contrary to popular belief, not all channels have a capacity. Most "reasonable" channels of interest do.

There are different and very interesting methods of proof of (1.8) and (1.9), but lack of space prevents our doing more than barely mentioning them. One method of proving (1.8) is based on the fact that if a code is chosen at random (!), by a reasonable and easily specified random process, the average error (of decoding) is very small. (This proves the existence of at least one code, and in a sense implies that "most" codes have a small probability of error!) In a second method of proving (1.8) the code is built up seriatim and arbitrarily until its prolongation is impossible; the code is then shown to have the desired length. (This again suggests that "most" codes are "good".) A third method involves a method of actually counting sequences. This last method can also be used to prove the strong converse. Another method of proving the strong converse essentially replaces counting sequences by measuring their volume. Finally the weak converse is proved by a simple and ingenious manipulation of entropies. Modifications and combinations of these methods are usually adapted to other channels. The proofs show up the combinatorial significance of the various entropies which occur in the statements and proofs of the theorems.

Consider now the dmc with the following difference: Suppose the sender can look over the receiver's shoulder and see what the latter is receiving. The sender can choose subsequent letters to be sent in order to correct previous reception, but he can communicate with the receiver only over the channel. The capacity of this channel is the same as if there were no feedback! This channel could occur if an earthy expedition landed on the moon. Naturally the power of the latter's transmission apparatus could not be great. The receiving station on earth, however, would have almost limitless power and could report with essentially perfect feedback the message actually received.

The term discrete is of engineering origin and really means finite. Channels which are not discrete have infinite input or output alphabets or both. The infinite alphabets may be countable or not. The usual method of treating such channels is to approximate their alphabets by finite alphabets. This is not always possible and often difficulties are encountered. When the alphabets are not denumerable measure-theoretic questions also arise.

Some references for this section:

[14], [21], [27], [28], [29], [40], [41], [45], [49], [54], [56], [61], [64], [83], [85], [86], [91], [94], [96], [98], [99], [101], [107], [112], [124], [128], [129], [132], [141], [143], [144], [160], [161], [162], [170], [173], [174], [175], [184].

2. Compound channels.

Consider now a dmc with this difference: Instead of a single cpf w there is given a set S^* of cpf's, say $S^{\mathfrak{S}} = \{w(\cdot/\cdot/s),\ s \in S\}$. Here the third index, s, distinguishes the cpf. The set S^* may have infinitely many elements. For each s, $w(j/i/s)$ is a cpf defined for $i = 1,\ldots,a$ and $j = 1,\ldots,b$. The compound channel transmits as follows: Each word of n letters (n-sequence) is transmitted according to some cpf in $S^{\mathfrak{S}}$; the cpf may vary arbitrarily in $S^{\mathfrak{S}}$ from one such word to another.

Let P_s now denote probability according to the cpf $w(\cdot/\cdot/s)$. A code (n,N,λ) for the compound channel is a system (1.2) with all the requirements, except that (1.3) is replaced by the stronger requirement

$$P_s\{v(u_i) \in A_i\} \geq 1 - \lambda, \quad i = 1,\ldots,N;\ s \in S. \tag{2.1}$$

Thus, even if Maxwell's demon tried maliciously to vary the cpf so as to make things as difficult as possible, the probability that any word sent would be incorrectly understood by the receiver is $\leq \lambda$.

The question is, how long can codes be and still meet this stronger requirement (2.1)? It must be borne in mind that the cpf's in $S^{\mathfrak{S}}$ may be very "antithetical" to each other. The fact is that theorems exactly like those for the dmc hold for the compound channel. Thus the maximum length of the code depends on a constant called (as in the case of dmc) the capacity (C_1 say) of the compound channel.

If C_1 were 0 in most cases, little could be done with a compound channel. Let C(s) be the capacity of the dmc with the

single cpf $w(\cdot|\cdot|s)$. Define

$$C_2 = \inf_{s\in S} C(s) = \inf_{s\in S} \max_{\pi} \{H(\pi'|s) - \sum_{i=1}^{a} \pi_i H[w(\cdot|i|s)]\}.$$

Then obviously we have $C_1 \leq C_2$, for the demon could use the "worst" cpf for every word, that is, the one with the smallest capacity. (If S is an infinite set and there is no worst cpf, one uses a cpf with a capacity arbitrarily close to the infimum.) The fact is that

$$C_1 = \max_{\pi} \inf_{s\in S} \{H(\pi'|s) - \sum_{i=1}^{a} \pi_i H[w(\cdot|i|s)]\},$$

and, surprisingly and pleasantly, C_1 is not 0 unless C_2 is 0. Thus C_1 is not 0 unless S^* contains a cpf whose capacity is 0 (or a sequence whose capacities approach 0).

Consider now a compound channel as above except that the receiver now knows which cpf is being used (but the sender does not). It has been shown that the capacity of this channel is also C_1. Thus knowledge of the cpf by the receiver alone does not increase the capacity!

Consider the compound channel as above, except that the cpf is now known to the sender but not to the receiver. The capacity is then C_2, which in general is greater than C_1.

In all of the above results S^* may contain infinitely many elements, indeed, non-denumerably infinitely many elements, and the cpf is chosen arbitrarily at the beginning of transmission of each word by the "jammer". Nevertheless, the fact that the same

cpf (although arbitrarily chosen) is used for every letter of the word has essentially the effect that, to a satisfactory approximation, the set $S^{\mathcal{C}}$ can be replaced by a finite set or at least one with $2^{\alpha\sqrt{n}}$ cpf's, where α is suitably chosen. This is always an essential step of the proof. Suppose now that the cpf varies arbitrarily from letter to letter of a word. The above reduction is now no longer possible, previously used methods no longer apply, and the problem becomes very difficult. Partial results have been proved in [84] and a complete solution announced without proof in [34] . Since a theorem announced in [34] is incompatible with a result proved in [84] it is clear that the channels treated are not the same.* While awaiting publication of the results announced in [34] and [35] one can repeat, without fear of contradiction, that the problems invovled in these "arbitrarily varying channels" are very difficult.

Suppose that the cpf varies arbitrarily from letter to letter, but with some limitations. For example, suppose that the number of changes from one cpf to another is not greater than a fixed multiple of n^{α}, $\alpha < 1$. In the latter case it is easy to prove that the problem can be reduced (and hence solved) to the (compound) case where the same cpf governs the transmission of each letter.

We spoke of the above problems as if neither the sender nor the receiver knew the cpf. Of course the problem of arbitrarily varying channels has been studied where either or both know the cpf for each letter. In fact, the capacity of the channel where both know the arbitrarily varying cpf is the smallest of the cpf's in the set $S^{\mathcal{C}}$.

*Randomized codes are used in [34] but are not admitted in [84] .

Perhaps this is the time briefly to mention the question of randomization. Conceivably the sender could use randomized encoding, i.e., each sender's message could be represented by a probability distribution over sequences in the input alphabet. After the sender has decided on the message he performs a chance experiment with the corresponding probability distribution and actually sends the resulting sequence. Randomized decoding is defined similarly in an obvious way. Randomized codes have been studied to some extent, and further studies are in process. Generally speaking, randomized decoding seems to provide little advantage, but under certain conditions randomized encoding actually helps by either making a longer code possible or by reducing the error. Indeed, the author of [34] states that his general results are valid only under randomized encoding. These results are for arbitrary varying channels, and an explanation of the need for randomized encoding may perhaps be the following. Suppose that there is a rational malevolent being, the "jammer" say, who chooses the arbitrarily varying cpf's so as to make communication between sender and receiver as difficult as possible. The utility of randomized encoding seems to be to protect the sender against the jammer. Even when the jammer knows the message to be sent, if he doesn't know the actual sequence which will represent it he may not be able to choose the sequence of cpf's which will most efficiently jam it. No such utility accrues to randomized decoding, and the sender can do best by voting for the message with the highest probability. (This is not strictly correct in the present model. The messages to be sent are chosen

arbitrarily and one cannot speak of the a posteriori probability of a message after the resulting chance sequence has been received. However, intuitively this is near enough, and it will be made precise in the next paragraph but one.) This discussion also points up the difference between two channels, each with arbitrarily varying cpf's (from letter to letter). In one channel the jammer knows the actual sequence being sent before it is sent, in the other he knows only the probability distribution over input sequences. Naturally the capacity of the second channel is not less than the capacity of the first channel. The information theory literature is sometimes not entirely mathematically precise and such distinctions are often made only implicitly. It is likely that [34] treats the second channel; [84] certainly treats the first.

The preceding remarks suggest that some problems in information theory should be viewed as zero-sum two-person games between the sender (and receiver) and the jammer. Indeed, the form of the capacities in the several forms of the compound (stationary) channel, i.e. max inf and inf max of the "information function" in the above definitions of C_1 and C_2, suggest a game-theoretic background to the problem. A number of writers have made more or less positive assertions about this, but no specific proof of any coding theorem or any other important fact by game-theoretic methods is available in the literature. If there is an essential, non-trivial, and meaningful connection between the two theories it would be very interesting and useful to establish it precisely; it might well lead to further results in coding theory. In several papers correlated encoding and decoding has been used. Here the sender, before transmitting any

message, chooses a code at random, communicates the result of his random experiment to the receiver, and then sends the message according to the code selected. This procedure is repeated at each message. It seems to the writers that this procedure cannot seriously be considered as reflecting anything remotely resembling actual communication. Surely it is vastly more complicated for the sender to transmit to the receiver the designation of the code which is the outcome of the chance experiment than it is to transmit the message itself. Yet a new code must be transmitted with each message! No doubt, however, problems involving correlated encoding and decoding have mathematical interest.

In most papers in information theory, especially those written by engineers, it is assumed that the message to be sent is itself chosen at random (usually with equal probability for each message). When this is so one can speak of the a posteriori probability of a message, after it has been passed through the channel and the chance sequence received. Naturally the average error is then minimized if the decoder decides that the message sent is the one with the largest a posteriori probability; this is called maximum likelihood decoding. Maximum likelihood decoding is simple and unambiguous only if there is only one cpf for the channel. If there is more than one cpf then different messages can have maximum a posteriori probability according to different cpf's and often a difficult theory is needed for a decision. Returning to the first and basic case studied, that of a dmc with a single cpf, the fact is that the two cases, that of average error with the message chosen by a (known or unknown) random mechanism, and that

of maximum error with messages chosen arbitrarily, are essentially the same. This is also true of many other channels. This fact contradicts the statement, made by some very great mathematicians and widely believed in engineering circles, that no theory is possible without knowing the statistics of the (randomly chosen) message. For example, the theory developed in Chapters 3 and 4 of [15] neither assumes the existence of a random mechanism for choosing messages nor makes any use of it.

A number of writers have stated, mostly without proof, that there is a basic and meaningful connection between information theory and the theory of statistical inference, and some of them have attempted formally to set up a theory which would exhibit this putative connection. It seems to us that to establish a basic and meaningful connection between two theories requires either that one obtain a common framework from which one can derive some of the basic theorems in both theories, or that one derive important old or new theorems in one theory by use of theorems or methods of the other. By this essential standard no meaningful connection between information theory and the theory of statistical inference has yet been established. Of course this does not prove that no such connection exists.

Some references for this section:

[3], [4], [11], [16], [17], [18], [19], [20], [23], [28], [29], [34], [35], [37], [47], [69], [70], [73], [76], [80], [81], [82], [84], [92], [95], [102], [104], [105], [123], [145], [154], [156], [157], [171], [176], [177], [178], [180], [182], [189], [194], [195], [198].

3. Error bounds. Sequential decoding.

Suppose given a dmc with capacity C. Let $0 < R < C$ and consider all codes of word length n and code length 2^{nR} for this channel. (In such a case one is said to be transmitting ar "rate" R.) It is not difficult to show that there exists two positive numbers, say D_1 and D_2, such that, among these codes, there exist one for which λ, the maximum error of decoding any word, satisfies $\lambda < D_1 \exp\{-nD_2\}$; this is summarized by saying that the error decays exponentially (with n). It is true for most channels, not only the dmc, and probably all channels of practical importance, that the error decays exponentially with n. The proof of this is usually quite simple and requires only an almost trivial modification of the proof of the coding theorem. An intuitive explanation is perhaps this: The codes we are considering are of such small length (approximately $2^{-n(C-R)}$ of the length they could have for a fixed λ) that there are great gaps among the different u_i's, and they can be distinguished (decoded) by the decoder with very great accuracy. Exponential decay of error is essentially due to the fact that the probability that the mean $n^{-1}\sum_1^n X_i$ of independent, identically distributed chance variables $X_1, X_2, \ldots, X_n$, shall exceed any fixed number larger than than their common expected value, decreases exponentially with n.

The school of electrical engineers working in information theory, whose intellectual center is the Massachusetts Institute of Technology, regards the determination of the best (i.e., largest possible) D_2 as one of the principal and most important problems of information theory. Determinations of bounds on D_1 is

considered of negligible importance. The reason given for the importance of the problem is that the complexity of the apparatus for coding and decoding goes up, roughly speaking, exponentially with n, so that it is important to know the smallest n for which one can achieve a desired rate R and a desired (usually small) upper bound on the error. Even for the dmc the problem is of formidable difficulty. Previous attempts consisted of using randomized coding theorem to get a lower bound on D_2 and sphere packing methods* to get an upper bound. It was thought that these two bounds agreed over a certain range of R, so that D_2 was determined for this range, but errors were found in the arguments. A recent new effort has succeeded in determining D_2 for part of the range of R. The argument is difficult and does not seem to lend itself to intuitive description or summary. At least that part of it which gives a lower bound on D_2 can be carried over with little change to many other channels. The value of D_2 for all R is as yet unknown, although approximations are available.

We now turn to another subject of major investigation among engineers, sequential decoding. This is one of the most beautiful of all ideas in coding theory and one of the most important for practical application. Unfortunately for the mathematician, it does not seem to lend itself to elegent mathematical theorems. Even an approximate description of the method would require essentially the reproduction of at least a short paper on the subject or the reproduction of the appropriate

* for a description of these methods see, e.g. [4] p. 227

chapter of a book. This is impossible for us, but perhaps the following lines will help to form some idea.

The actual application of the codes hitherto discussed would always occur in connection with a computer. The code would be stored in the computer and the latter would be indispensable in both encoding and decoding; the latter process requires many more computations than the first. The volume of material to be stored and the number of operations to be performed increase exponentially with n and soon exceed the capacity of even very large modern computers. This raises the problem of finding methods which can be carried out practically. Sequential decoding is intended to be such a method.

We pause for a moment for an intuitive discussion of what makes efficient coding. If the transmission of any one letter is repeated a sufficient number of times then, except in certain obvious special cases, the decoder (receiver) can identify the letter being sent with a probability as close to one as desired. In this way any desired message could be transmitted with any desired degree of accuracy. The trouble with this naive method is that it is grossly inefficient; in terms of our previous parameters, for given λ and R an enormous n is generally required. What makes for efficient decoding are the differences between entire words rather than between individual letters; the letters of a word reinforce each other, so to speak, so that even if several letters are misunderstood the entire pattern still remains clear. This is called "redundancy," as distinct from simple repetition. For a homely example, consider the problem of reading

every letter of a manuscript written in poor handwriting. If the reader is familiar with the subject or even the language he can often reconstruct illegible letters or words from the context. This is impossible if what is written is made up of nonsense syllables or material in a completely unknown language. (Although in the latter case one can start looking for patterns (i.e., redundant elements) as crypto-analysts do.) The idea of sequential decoding is to introduce redundancy into the decoding of individual letters, while avoiding the construction of codes which require the storage of, and calculation with, exponentially many sequences. We should emphasize, however, that there is not just one method of sequential decoding, but a number of variations. We shall describe a typical, but by no means unique, method.

In the basic and simplest description of sequential decoding it is assumed that the channel is binary symmetric and that one has the problem of reproducing a stream (doubly-infinite sequence) of chance "information" digits which take the values 0 and 1 with equal probability, all independently of each other. (A binary symmetric channel reproduces each of the two digits correctly with probability 1-p, say, and reverses the digits with probability p.) Suppose that the digits actually realized are $\ldots, m_{-1}, m_0, m_1, \ldots$. Let m and k be integers, and suppose that the rate $R = \frac{1}{m}$, and that mk is the "constraint length". Each information digit will be coded into m digits which are then transmitted over the channel. Each of these m digits is a linear combination of

the information digit being sent, say m_0, and its (k-1) immediate predecessors, $m_{-1},\ldots,m_{-(k-1)}$. Hence the "effect" of any information digit extends over mk digits, transmitted and received, and one does not decode this information digit until mk digits have been received; this delay is a price paid for using the method. The decoder is now supposed to know the preceding (k-1) information digits. (He has decoded them correctly with very high probability.) He begins a search which will end in decoding the current information digit. This search is impossible to describe under our present limitations. It is based on the fact that, with very high probability (depending upon mk) the "distance" (this depends on the particular sequential decoding procedure) between the received sequence of mk digits and a transmitted sequence of mk digits which corresponds to <u>any</u> information sequence $\bar{m}_0, m_1',\ldots,m_{k-1}'$, where $\bar{m}_0$ is the digit different from m_0, is large. By good sequential decoding procedures one can relatively quickly eliminate as possiblities all sequences which start with $\bar{m}_0$, with a small probability of error. Here "relatively quickly" refers to the average number of required searches and computations as compared to the probability of error. The above description is very ~~very~~ crude and incomplete, as any description of this brevity must be, and at best can only give an inkling of the flavor of this beautiful idea.

The published results in the literature of sequential decoding consist of descriptions of different schemes, and the

theorems are statements about the expected number of computations needed by the scheme under certain conditions and the probability of error in decoding. These results are often clever and ingenious and considerable difficulties have to be surmounted in obtaining them. Unfortunately for the mathematicians, the theorems are almost never elegant, as are the theorems in the Shannon theory. There are no clear cut proofs that a certain procedure is optimal. Perhaps these will still come.

We close this section with a brief description of a totally different and also very clever idea. Consider a channel with feedback where independent, identically distributed Gaussian errors are added to each transmitted signal, and $\frac{1}{n}$ times the sum of the squares of the n signals which represent a word is bounded above by a given constant (the "average power"). According to this idea the message is coded in an arbitrary but fixed manner into one of a set of equally spaced points of an interval, and this point (number) is transmitted; this is the first of the sequence of n signals which will be used to transmit the message. The i^{th} transmitted signal, $i = 2,\ldots,n$, is a suitably chosen linear function of the message and all previously received signals. The decoding of the message after the n^{th} received signal is also very simple: one decodes the message sent as the one corresponding to that one of the equally spaced points which is nearest to the n^{th} received signal. It has been proved that this method is optimal in a very natural and reasonable sense, and it is clear

from the description that it involves a minimum of encoding and decoding computations. Unfortunately, it possesses one very serious drawback. If n is sufficiently large the probability is very close to one that at least one of the signals will require for its transmission an amount of power (i.e., the square of the signal) which exceeds any fixed bound (and hence the "capacity" of the instrument).

Some references for this section:

[8] [9], [15], [17], [32], [33], [43], [44], [46], [52], [53], [54], [58], [68], [77], [79], [87], [88], [93], [96], [97], [121], [126], [127], [135], [136], [137], [139], [140], [147], [150], [155], [158], [159], [163], [183], [186], [187], [190], [192], [197], [199], [200], [201], [202], [203].

4. Coding with a fidelity criterion. - Work of the Russian school.

Shannon [148] and other writers have studied the following problem: An (infinite) sequence of information digits, i.e., values taken by a sequence of chance variables with a known distribution (the chance variables are usually independently and identically distributed, but this is not essential) are produced by a source. After the source has produced n digits the latter are coded into a sequence of n' digits; this sequence is transmitted over a noisy channel, and the received n' - sequence is decoded into a sequence of n information digits. (The actual formulation is slightly more general).

There is a "fidelity criterion" which measures the 'distortion' between the sequence of n digits thus decoded and the sequence of n digits produced by the source. The results obtained in the theory deal with such questions as the minimum ratio $\frac{n'}{n}$ needed so that the distortion not exceed a given bound, and the geometric problem of the minimum number of n-sequences of information digits needed to 'span' the space of such sequences of information digits to within a specified bound on the distortion. We shall not describe these results. Instead, we shall describe the generalization of the above model whose study has been a major occupation of the Russian school of information theorists, and shall describe a typical and important result of the Russian school about this model.

Let $(\xi', \eta'), \ldots, (\xi^t, \eta^t) \ldots$ be a sequence of chance variables, the pair (ξ^t, η^t) being defined on the probability space (Ω^t, B^t, P^t) with values in the measure spaces (X^t, S_X^t) and (Y^t, S_Y^t), respectively. The information of this pair (ξ^t, η^t) relative to each other is defined by

$$I(\xi^t, \eta^t) = \int_{X^t \times Y^t} \log f^t_{\xi\eta}(x, y)\, P^t_{\xi\eta}(dx, dy)$$

where $P^t_{\xi\eta}$ is the joint probability distribution of (ξ^t, η^t) determined by P^t, and $f^t_{\xi\eta}$ is the density of $P^t_{\xi\eta}$ with respect to $P^t_\xi \times P^t_\eta$, the product of the marginal measures. It is always supposed that the integral in I is finite, so that $f^t_{\xi\eta}$ is finite with probability one. One denotes $\log f^t_{\xi\eta}$ by $i^t_{\xi\eta}$ and calls it the information density. The sequence (ξ^t, η^t) is called information stable if, for all t sufficiently large, $I^t(\xi, \eta) > 0$ and

$$\frac{i^t_{\xi\eta}(\xi, \eta)}{I(\xi^t, \eta^t)}$$

converges stochastically to one.

Let (X, S_X) and (Y, S_Y) be two measure spaces, respectively the space of input messages and output messages. Let W be a given set of distributions $P_{x,y}$ of chance variables x, y with values

in $X \times \tilde{X}$, such that the marginal distributions of ξ are all the same. This set is called the message $\{W\}$, and p_ξ is called the distribution of the input message. Any pair $(\xi, \tilde{\xi})$ of chance variables whose joint distribution belongs to W is said to satisfy the conditions of reproduction W. Without loss of much generality one limits one's self to W defined as follows: Suppose given real functions $\rho_i(x,\tilde{x})$, $i = 1,\ldots, M$, defined on $X \times \tilde{X}$ and measurable with respect to the σ-algebra $S_X \times S_{\tilde{X}}$, and an M-dimensional set $\tilde{W}$. W consists of all distributions $p_{\xi\tilde{\xi}}$ (with the same marginal distribution of ξ) for which the vector whose i^{th} component, $i = 1,\ldots,M$, is

$$E\rho_i(\xi, \tilde{\xi}),$$

belongs to $\tilde{W}$. The message entropy with accuracy of reproduction W^* is defined to be

$$H(W) = \inf_W I(\xi, \tilde{\xi}).$$

The sequence of messages $\{W^t\}$ is called information stable if there exists an information stable sequence of pairs $(\xi^t, \tilde{\xi}^t)$ such that the t^{th} pair satisfies the condition of reproduction W^t and

$$\lim \frac{I(\xi^t, \tilde{\xi}^t)}{H(\xi^t)} = 1.$$

The problem of obtaining general sufficient conditions for the information stability of a sequence of messages is bound up with the problem of obtaining sufficiently general conditions for the information stability of a sequence of pairs of chance variables.

Let (Y, S_Y) and $(\tilde{Y}, S_{\tilde{Y}})$ be two measure spaces which serve as the space of input signals and space of output signals, respectively. Let $Q(y, \tilde{A})$, $y \in Y$, $\tilde{A} \in S_{\tilde{Y}}$, be a transition function such that a) for fixed y, $Q(y, \quad .)$ is a probability measure on $S_{\tilde{Y}}$ b) for fixed $\tilde{A} \in S_{\tilde{Y}}$, $Q(\ ., \quad \tilde{A})$ is measurable with respect to the σ-algebra S_Y. Let V be a given set of probability distributions on $(Y \times \tilde{Y}, S_Y \times S_{\tilde{Y}})$. The system consisting of Y, $\tilde{Y}$, Q, and V is called "the transmitter" and will be denoted for brevity by $\{Q, V\}$. The chance variables $\eta, \tilde{\eta}$ with values in Y and $\tilde{Y}$, respectively, are "connected by the transmitter $\{Q, V\}$" if their joint distribution belongs to V and for any $\tilde{A} \in S_{\tilde{Y}}$ the conditional probability

$$P\{\tilde{\eta} \in \tilde{A} \mid \eta\} = Q(\eta, \tilde{A})$$

with probability one. Again, without loss of much generality, one limits one's self to sets V defined as follows: Suppose given real functions $H_i(y, \tilde{y})$, $i = 1, \ldots, N$, defined on $Y \times \tilde{Y}$, and an N-dimensional set $\tilde{V}$. The set V of distributions consists of all distributions of the pair $(\eta, \tilde{\eta})$ such that the

vector whose i^{th} component, $i = 1, \ldots, N$, is

$$E\ \pi_i(\eta, \tilde{\eta}),$$

is in $\bar{V}$. When the π_i depend only on y the constraint imposed by V is on the input signal only. The capacity of the transmitter $\{Q,V\}$ is defined to be

$$C\ (Q,V) = \sup_V I\ (\eta, \tilde{\eta}).$$

The sequence of transmitters $\{Q^t, V^t\}$ is said to be information stable if there exists an information stable sequence of pairs of chance variables $(\eta^t, \tilde{\eta}^t)$ such that the t^{th} pair is connected by the t^{th} transmitter and such that

$$\lim \frac{I(\eta^t, \tilde{\eta}^t)}{C(Q^t, V^t)} = 1$$

All published results concern themselves only with information stable sequences of transmitters.

The message $\{W\}$ is said to be transmissible by means of the transmitter $\{Q,V\}$ if there exists a sequence of four chance variables $(\xi, \eta, \tilde{\eta}, \tilde{\xi})$ such that: a) this sequence is a Markov chain b) the pair $(\Xi, \tilde{\Xi})$ satisfies the conditions of reproduction W c) the pair $(\eta, \tilde{\eta})$ is connected by the transmitter $\{Q,V\}$. The intuitive meaning of the above is as follows:

The input message is the chance variable ξ with given distribution p_ξ. The input message ξ is coded into the input signal η, which is sent over the transmitter (channel) and received as $\tilde{\eta}$. Then $\tilde{\eta}$ is decoded into the output message $\tilde{\xi}$. The transmitter is given and W represents the desired accuracy of reproduction. The conditional probability of η, given ξ, is the randomized encoding procedure. The conditional distribution of η, given η (i.e., $\varphi(\eta, \cdot)$) is the distribution of the received signal. The conditional probability of $\tilde{\xi}$, given $\tilde{\eta}$, is the randomized decoding procedure.

It is easy to prove that a necessary condition that the message $\{W\}$ be transmissible by means of the transmitter $\{Q,V\}$ is that

$$H(W) \leq C(Q,V).$$

A principal concern of the writers of the Russian school is to prove that, asymptotically and under additional reasonable regularity conditions, this condition is also sufficient. We shall now describe a typical and important result. Let $\{W^t\}$ be a given sequence of messages and $\{Q^t, V^t\}$ a given sequence of transmitters. We define the distance $r(a,b)$ between two points of the same Euclidean space as the maximum absolute deviation between corresponding components of a and b. If U is a set in a Euclidean space let $[U]_\varepsilon$ denote the set of all

points within an r-distance of at most ϵ from some point of U. We now replace $\bar{W}$ by $[\bar{W}]_\epsilon$, and call the corresponding message $\{W_\epsilon\}$. We also replace $\bar{V}$ by $[V]_\epsilon$ and call the corresponding transmitter $\{Q, V_\epsilon\}$. We say that the message $\{W\}$ is transmissible by means of the transmitter $\{Q,V\}$ within an event of probability ϵ, if there exist four chance variables $\xi, \eta, \tilde{\eta}, \tilde{\xi}$ and a fifth chance variable $\tilde{\xi}'$, defined on the same space as $\tilde{\xi}$, such that

a) $(\xi, \eta, \tilde{\eta}, \tilde{\xi})$ form a Markov chain b) $(\xi, \tilde{\xi}')$ satisfy the conditions of reproduction W c) the pair $(\eta, \tilde{\eta})$ is connected by the transmitter $\{Q,V\}$ d) the probability that $\tilde{\xi} \neq \tilde{\xi}'$ is not greater than ϵ. Now let $\{W^t\}$ be a given sequence of messages and $\{Q^t, V^t\}$ a given sequence of transmitters, such that

a) $\lim H(W^t) = \infty$ b) $\lim \frac{H(W^t)}{C(Q^t, V^t)} < 1$

c) the number M^t of functions φ_i^t in the definition of the message and the number N^t of functions π_i^t in the definition of the transmitter are such that, for every $a > 0$,

$$M^t = o\left(\exp_2 \left\{ a\, H(W^t) \right\}\right)$$

and

$$N^t = o\left(\exp_2 \left\{ a\, C(Q^t, V^t) \right\}\right)$$

d) the sequence of transmitters $\{Q^t, V^t\}$ is information stable

e) for some $\{(\eta^t, \tilde{\eta}^t)\}$, a sequence of pairs with respect to which the sequence $\{Q^t, V^t\}$ is information stable, for some $b > 0$ and for every $a > 0$,

$$\max_{k=1,\dots,N^t} E\left\{\left|\eta_k^t(\xi^t,\tilde{\xi}^t)-E\eta_k^t(\xi,\tilde{\xi})\right|^{1+\delta}\right\}$$

$$= o(\exp_2\{a\,C(Q^t,V^t)\})$$

f) the sequence $\{W^t\}$ is information stable, and g) for some sequence $(\xi^t,\tilde{\xi}^t)$ of information stable sequences of chance variables with respect to which $\{W^t\}$ is information stable and which also satisfy the conditions of reproduction $\{W^t\}$, for some $b>0$ and every $a>0$,

$$\max_{k=1,\dots,M^t} E\left\{\left|\rho_k^t(\xi,\tilde{\xi})-E\rho_k^t(\xi,\tilde{\xi})\right|^{1+b}\right\} = o(\exp_2\{aH(W^t)\}).$$

Then, for every $\varepsilon>0$ there exists a number T such that for $t\geq T$ the message $\{W_c^t\}$ is transmissible by the transmitter $\{Q^t, V_\varepsilon^t\}$ "within an event of probability ε." Under additional conditions one can eliminate the phrase in quotation marks. One such set of conditions is that each of the sequences $\{M^t\}$, $\{N^t\}$, and

$$\sup_k \sup_{x,\tilde{x}} |\rho_k^t(x,\tilde{x})|$$

and

$$\sup_k \sup_{y,\tilde{y}} |\pi_k^t(y,\tilde{y})|$$

should be bounded.

Some references to this section:

[28], [29], [30], [34], [35], [38], [57], [61], [62], [64], [89], [92], [108], [112], [116], [117], [118], [119], [120], [122], [131], [132], [148], [152], [181], [196],

and also I and $I^{\frac{1}{2}}$

a) BOOKS

1 Abramson, N.M. (1963) Information Theory and Coding, McGraw Hill, New York.

2 Ash, R. (1965) Information Theory, Interscience Publishers, New York.

3 Bell, D.A. (1962) Information Theory and its Engineering Application, 3rd edition, Pitman, London.

4 Fano, R.M. (1961) Transmission of Information, M.I.T. Press, Cambridge, Mass., and John Wiley, New York.

5 Feinstein, A. (1958) Foundations of Information Theory, McGraw Hill, New York.

6 Fey, P. (1963) Informationstheorie, Akademie-Verlag, Berlin.

7 Fleishman, B.S. (1963) Constructive Methods of Optimal Coding in Noisy Channels, Akad. Nauk SSSR, Moscow.

8 Harman, W.W. (1963) Principles of the Statistical Theory of Communication, McGraw Hill, New York.

9 Helstrom, C.W. (1960) Statistical Theory of Information, Pergmon Press, New York.

10 Jelinek, F. (1968) Probabilistic Methods of Information Theory, McGraw Hill, New York.

11 Khinchin, A.I. Mathematical Foundations of Information Theory, 1957, Dover Publishers, New York (English translation).

12 Pinsker, M.S. Information and Stability of Random Variables and Processes, 1964, Holden-Day, San Francisco (English Translation).

13 Reza, F. (1961) An Introduction to Information Theory, McGraw Hill, New York.

14 Shannon, C.E. and Weaver, W. (1949) The Mathematical Theory of Communication, University of Illinois Press, Urbana, Illinois.

15 Wolfowitz, J. (1961), (1964) Coding Theorems of Information Theory (1961), Springer-Verlag, Berlin and Prentice-Hall, Englewood Cliffs, New Jersey. Second edition (1964), Springer-Verlag, New York.

16 Woodward, P.M. (1964) Probability and Information Theory, with Application to Radar, Pergamon Press, Oxford, New York.

17 Wozencraft, J.M. and Jacobs, I.M. (1965) Principles of Communication Engineering, J. Wiley and Sons, New York.

18 Wozencraft, J.M. and Reiffen, B. (1961) Sequential Decoding, M.I.T. Press and J. Wiley, New York.

19 Yaglom, A.M. and Yaglom, I.M. (1960) Probability and Information.

c) SURVEY PAPERS

I Dobrushin, R.L. (1961) Mathematical Problems in Shannon Theory of Optimal Coding of Information, Proc. Fourth Berkeley Sympos. Vol. I, 211-252.

II Dobrushin, R.L. (1963) Development of Information Theory in the USSR, Part III, Coding Theory, Investigation of Probability of Error for Optimal Transmission Methods, Izvest. Akad. Nauk, Techn. Kibernet., 5, 81-84.

III Elias, P. (1961) Progress in Information Theory in the U.S.A., 1957-1960, Part 1: Information Theory and Coding, IRE Transact. on Information Theory, 7, 128-131.

IV Gallager, R.G. (1964) Information Theory, Chapter IV in the Mathematics of Physics and Chemistry, Vol. II, H. Margenau and G.M. Murphy, ed., Van Nostrand, Princeton, New Jersey.

V Kotz, S. (1966) Recent Results in Information Theory, J. of Applied Prob. 3, 1-93. (Also a Matheun monograph (1966) under the same title).

VI Peterson, W.W. and Massey, J. (1963) Report on Progress in Information Theory in the U.S.A., 1960-1963, Part 2; Coding Theory, IEEE Transact. Information Theory 9, 4, 223-229.

VII Pinsker, M.S. (1963) Development of Information Theory in the USSR, Part II, Mathematical Foundations of Theory of Information Transmission, Izvest. Akad. Nauk, Techn. Kibernet., 5, 79-81.

VIII Pinsker, M.S. (1964) Mathematical Foundations of Theory of Optimal Information Coding, in Itogi Nauki; Math. Analysis, Probability Theory and Control, 1962, Institute of Scientific Information, Moscow, 197-210.

IX Savage, J.E. (1968) Progress in Sequential Decoding, in Advances in Communications Systems, Vol. 3, A.V.Balakrishnan, ed., Academic Press, New York, 149-204.

X Siforov, V.I. (1967), Problems in Information Transmission and Processing (Survey), Izvest. Akad. Nauk, Techn. Kibernet. 1, 76-81.

XI Siforov, V.I. and Tsybakov, B.S. (1963) Development of Information Theory in the USSR, Part I, General Review, Izvest. Akad. Nauk, Techn. Kibernet. 5, 74-78.

XII Thomasian, A.J. (1963) Report on Progress in Information Theory in the U.S.A., 1960-1963, Part 1; Foundations in Information Theory, IEEE Transact. Information Theory, 9, 4, 221-223.

c) Research Papers

1 Abramson, N. (1960) A partial ordering for binary channels. IRE Trans. Information Theory 6, 529-539.

2 Ahlswede, R. (1968) Beiträge zur Shannonschen Informationstheorie im Falle nichstationärer Kanäle, Z. Wahrscheinlichkeitsth. 10, 1, 1-42.

3 _____ (1969) The weak capacity of averaged channels, Z. Wahrscheinlichkeitsth., 11,

4 _____ (1968) Certain results in coding theory for compound channels. I, to appear in Proc. of Colloquium on Inform. Theory, Debrecen, 1967.

5 Ash, R. B. (1963) Capacity and error for a time-continuous Gaussian channel, Information and Control 6, 14-27.

6 _____ (1964) Further discussion of a time-continuous Gaussian channel. Information and Control 7, 78-83.

7 Augustin, U. (1966) Gedächtnisfreie Kanäle für diskrete Zeit, Z. Wahrscheinlichkeitsth. 6, 10-61.

8 _____ (1968) Error estimates for low rate codes, Manuscript, Cornell University.

9 _____ (1968) On the second order estimates for the coding theorem and its strong converse, Manuscript, Cornell University.

10 _____ (1968) The capacity of the product channel, Manuscript, Cornell University.

11 Birch, John J. (1963) On information rates for finite-state channels. Information and Control 6, 372-380.

12 Blachman, N. M. (1962) On the capacity of a band-limited channel perturbed by statistically dependent interference. IRE Trans. Information Theory 8, 1, 48-55.

13 _____ (1962) The effect of statistically dependent interference upon channel capacity. Ibid. 8, 5, 53-57.

14 Blackwell, D. (1960) Infinite codes for memoryless channels. Ann. Math. Statist. 30, 1242-1244.

15 _____ (1961) Exponential error bounds for finite-state channels. Proc. Fourth Berkeley Symposium on Mathematical Statistics and Probability 1, 57-63. University of California Press, Berkeley.

16 Blackwell, D., Breiman, L., Thomasian, A. J. (1958) Proof of Shannon's transmission theorem for finite-state indecomposable channels. Ann. Math. Statist. 29, 1209-1220.

17 _____, _____, _____ (1959) The capacity of a class of channels Ann. Math. Statist. 30, 1229-1241.

18 _____, _____, _____ (1960) The capacities of certain channel classes under random coding. Ann. Math. Statist. 31, 558-567.

19 Breiman, L. (1960) On achieving channel capacity in finite-memory channels. Illinois J. Math. 4, 246-252.

20 _____ (1960) Finite-state channels, Trans. 2nd Prague Conf. Information Theory, Statist. Decision Functions, Random Processes, 49-60, Prague.

21 Campbell, L. L. (1965) A coding theorem and Rényi entropy, Information and Control 8, 423-429.

22 _____ (1966) Definition of entropy by means of a coding problem. Z. Wahrscheinlichkeitsth. 6, 113-118.

23 Carlyle, J. W. (1964) On the external probability structure of finite-state channels. Information and Control 7, 385-397.

24 Chang, T. T., and Lawton, J. G. (1962) Partial ordering of discrete channels. IRE Internat. Convention Record 10, 190-199.

25 _____, _____ (1964) On the comparison of communication channels. IRE Trans. Information Theory 10, 97-98.

26 Cherny, J. (1966) Information transmission in the case of coding by finite automata, Kibernetika, (Prague), 2, 5, 397-415.

27 Dobrushin, R. L. (1958) The transmission of information along a channel with feedback, Teor. Veroyatnost. i Primenen., 3, 395-412 (in Russ

28 _____ (1959) A general formulation of the fundamental Shannon theorem in information theory. Uspehi Mat. Nauk. 14, 3-104 (in Russian).

29 _____ (1959) General formulation of Shannon's basic theorems of the theory of information. Dokl. Akad. Nauk. SSSR 126, 474-479 (in Russian

30 _____ (1959) The optimal transmission of information along a channel with unknown parameters, Radiot. i Elektron., 4, 1951-1956 (in Russian).

31 _____ (1962) Optimal binary codes for small rates of transmission of information. Teor. Verojatnost. i Primenen. 7, 208-213 (in Russian).

32 _____ (1962) Asymptotic bounds for the probability of error of information transmission over a discrete memoryless channel with a symmetric transition matrix. Teor. Verojatnost. i Primenen. 7, 283-311 (in Russian).

_____(1962) An asymptotic bound for the probability of error of information transmission through a channel without memory using feedback. Problemy Kibernet. 8, 161-168 (in Russian).

_____(1963) Unified methods of information transmission for discrete cnannels without memory and for communications with independent components. Dokl. Akad. Nauk SSSR 148, 1245-1248 (in Russian).

_____(1963) Unifoed methods of information transmission. The general case. Dokl. Akad. Nauk SSSR, 149, 16-19 (in Russian).

_____(1964) On the sequential method of Wozencraft-Reiffen. Problemy Kibernet. 12, 113-123 (in Russian).

_____(1967) Shannon's theorems for channels with errors in synchronization, Probl. Pered. Inform., 3, 4, 18-36 (in Russian)..

Dobrushin, R.L. and Tsybakov, B.S. (1962) Information Transmission with additional noise. IRE Trans. Information Theory 5, 293-304.

Drygas, H. (1965) Verschlusselungstheorie fur symmetrische Kanale, Z. Wahrscheinlichkeitsth. 4, 121-143.

Eisenberg, E. (1963) On channel capacity. Internal Technical Memorandum M-3, Electronic Research Labs., U. of California.

Elias, P. (1955) Coding for noisy channels, IRE Convention Record, No. 4, 37-46.

_____(1955) Coding for two noisy channels, Proc. Third London Symp. on Inform. Theory, 61-76. Butterworths Scientific Publications, London.

_____(1963) Information theory and decoding computations. Proc. Symp. Appl. Math. XV, 51-58, Amer. Math. Soc., Providence, R.I.

44 Fano, R. M. (1963) A heuristic discussion of probabilistic decoding. IEEE Trans. Information Theory 9, 64-74.

45 Feinstein, A. (1954) "A new basic theorem of information theory", Trans. IRE, PG-IT, 2-22.

46 _____ (1955) "Error bounds in noisy channels without memory," IRE Trans. Inf. Theory, 1, 13-14.

47 _____ (1959) On the coding theorem, and its converse for finite-memory channels. Information and Control 2, 25-44.

48 Fleischer, J. (1958) The central concepts of communication theory for infinite alphabets, J. Math. Phys., 37, 223-228.

49 Fleishman, B. S. (1963) Fundamental theorems of the constructive theory of optimal coding for a discrete noisy channel, Radiotehn. i. Elektron. 8, 1291-1300.

50 _____ (1965). Parallel decoding, Dokl. Akad. Nauk SSSR, 163 6, 1331-1333.

51 _____ (1966) Parallel decoding and optimization of data processing of space experiments, in Space Research, Vol. 6, Spartan Books, Washington, D. C., 269-279.

52 Forney, D. G. (1968) Exponential error bounds for erasure, list, and decision feedback schemes, IEEE Transact. Inform. Theory, 14, 2, 206-220.

53 Gallager, R. G. (1965) A simple derivation of the coding theorem and some applications. IEEE Trans. Information Theory, 11, 3-17.

64 Gallager, R. G.(1964) Bounds on error probability for a class of channels with memory, Intern. Conf. Microwaves, Circuit Theory and Inform. Theory, Tokyo, Part 3, 21-22.

55 _____ (1966). Channel capacity and coding for additive Gaussian noise channels with power and frequency constraints (Abstract) IEEE Trans. Inform. Theory, 12, 2, 273.

56 Gelfand, I. M., Kolmogorov, A. N. and Yaglom, A. M. (1956) Towards a general definition of the quantity of information. Dokl. Akad. Nauk SSSR, 11, 745-748 (in Russian).

57 Goblick, T. J. (1962) Coding for a discrete information source with a distortion measure. Ph. D. thesis, M.I.T.

58 Goutmann, M. M. (1967) Chernoff bound for channels with infinite memory, IEEE Transact. Inf. Theory, 13, 463-467.

59 Holsinger, J. L. (1965) Digital communication over fixed time dispersive channels, 1st IEEE Annual Communs. Conven. Boulder, Colo., 1, 731-735.

60 Helgert, H. J. (1967) A partial ordering of discrete, memoryless channels, IEEE Transact. Inf. Theory, 13, 360-365.

61 Hu, Go-Din (1961) Three kinds of converses to Shannon's theorem in information theory. Acta Math. Sinica 11, (in Chinese); translated as Chinese Math. 2, (1963), 293-332.

62 _____ (1962) Information stability of sequences of channels. Teor. Verojatnost. i Primenen. 7, 271-282 (in Russian).

63 _____ (1962) On information quantity. Teor. Verojatnost. i Primenen. 7, 447-455 (in Russian).

64 _____ (1964) On Shannon theorem and its converse for sequences of communication schemes in the case of abstract random variables. Trans Third Prague Conf. Information Theory, Statist. Decision Functions and

65 Hu, G o-din ; Shen-Shi-yi (1965) Some coding theorems for almost-periodic channels, Acta. Math. Sinica 15, 136-152 (Chinese); transl. as Chinese Math. - Acta 6, (1965) 437-455.

66 Huang, R. Y. and Johnson, R. A. (1962) Information capacity of time-continuous channels, IRE Trans. Information Theory 8, 191-198.

67 _____ (1963) Information transmission with time-continuous random processes. IEEE Trans. Information Theory 9, 84-89.

68 Jacobs, I. M. and Berlekamp, E. R. (1967) A lower bound to the distribution of computation for sequential decoding, IEEE Trans. Inf. Theory, 13, 167-174.

69 Jacobs, K. (1959) Die Übertragung diskreter Informationen durch periodische und fastperiodische Kanäle, Math. Ann. 137, 125-135.

70 _____ (1960) Über die Durchlasskapazitat periodischer und fastperiodischer Kanäle. Trans. 2nd Prague Conf. Information Theory, 231-249. Publ. House Czechoslovak Akad. Sci., Prague.

71 _____ (1962) Über Kanäle vom Dichtetypus. Math. Z. 78, 151-170.

72 _____ (1962) Almost periodic channels. Colloquium on combinatorial methods in probability theory. Matematisk Institut, Aarhus University, 118-136.

73 _____ (1966) Almost periodic sources and channels, Z. Wahrscheinlichkeitschth. 9, 65-84.

74 Jelinek, F. (1963) Loss in information transmission through two-way channel. Information and Control 6, 337-371.

75 _____ (1964) Coding for and decomposition of two-way channels. IEEE Trans. Information Theory 10, 5-17.

76 _____ (1965) Indecomposable channels with side information at the transmitter. Information and Control 8, 36-55.

77 _____ (1968) Evaluation of expurgated bound exponents, IEEE Transact. Inform. Theory; 14, 3, 501-505.

78 Karmazin, M. A. (1964) Solution of a problem of Shannon. Problemy Kibernet. 11, 263-266 (in Russian).

79 Kashyap, R. L. (1968) Feedback coding schemes for an additive noise channel with a noisy feedback link, IEEE Transact. Inform. Theory, 14, 3, 471-480.

80 Kellog, P. J. and D. J. Kellog (1954) Entropy of information and the odd ball problem, J. Appl. Phys., 25, 1438-1439.

81 Kelly, J. L. Jr. (1956) A new interpretation of information rate, Bell System Tech. J., 35, 917-926.

82 Kesten, H. (1961) Some remarks on the capacity of compound channels in the semicontinuous case. Information and Control 4, 169-184.

83 Khinchin, A. I., On the basic theorems of information theory, Uspehi Mat. Nauk, 11 (1956), 17-75 (in Russian).

84 Kiefer, J. and Wolfowitz, J. (1962) Channels with arbitrary varying channel probability function. Information and Control 5, 44-54.

85 Kolmogorov, A. N. (1957) The theory of the transmission of information, 1956. Plenary session of the Academy of Sciences of the USSR on the automatization of production, Moscow, Izd. Akad. Nauk SSSR. 66-99 (in Russian).

86 Kolmogorov, A. N. (1965) Three approaches to the definition of the concept "quantity of information", Probl. Pered. Inform., 1, 1, 3-11 (in Russian).

87 Koshelev, V. N. (1966) An estimate on the complexity of sequential decoding in the case of random tree-codes, Probl. Pered. Inform. 2, 2, 12-28 (in Russian).

88 Kotz, S. (1961) Exponential bounds on the probability of error for a discrete memoryless channel. Ann. Math. Statist. 32, 577-582.

89 Libkind, L. M. (1965) ε-entropy of discrete information sources, Probl. Pered. Inform., 1, 3, 48-55.

90 Libkind, L. M. (1967) Two-way discrete channels without memory, Probl. Pered. Inform., 3, 2, 37-46.

91 Lomnizky, Z. A. and S. K. Zaremba (1959) The asymptotic distributions of the amount of transmitted information, Information and Control, 2, 266-284.

92 Ma Hsi-Wen (1964) Feinstein's lemma for a finite system of channels. Chinese Math. 5, 316-329.

93 Mao Shi-Sun (1965) Asymptotic of the optimal probability of error for the transmission of information in the channel without memory which is symmetrical of pairs of input symbols for small rates of transmission. Teor. Verojatnost. i Primenen. 10, 167-175 (in Russian).

94 Meister, B. and Oettli, W. (1967) On the capacity of a discrete, constant channel, Inform. and Control, 11, 341-351.

95 Metzner, J. J. (1965) An interesting property of some infinite-state channels, IEEE Transact. Inform. Theory, 11, 310-312.

96 Molchanov, S. A. (1967) Transmission of a finite number of messages in a binary asymmetrical channel, Probl. Pered. Inform., 3, 10-17 (in Russian).

97 Morozov, V. A. (1967) On sequential decoding with an optimal probability of error, Izvest. Akad. Nauk, Tehnich. Kibernetika, 3, 164-167.

98 Muroga, S. (1953) On the capacity of a discrete channel I. J. Phys. Soc. Japan 8, 484-494.

99 _____ (1956) On the capacity of a discrete channel II. J. Phys. Soc. Japan 11, 1109-1120.

100 _____, "On the capacity of a noisy continuous channel," IRE Trans. Inf. Theory, IT-3, (1957), 44-51.

101 Nedoma, J. (1957) The capacity of a discrete channel. Trans. First Prague Conf. Information Theory, Statistical Decision Functions and Random Processes, 143-182. Prague.

102 _____ (1960) On non-ergodic channels. Trans. 2nd Prague Conf. Information Theory, 363-395. Prague.

103 _____ (1964) The synchronization for ergodic channels. Trans. 3rd Prague Conf. Information Theory, Statist. Decision Functions and Random Processes, 529-539. Prague.

104 _____ (1963) Die Kapazität der periodischen Kanäle. Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 2, 98-110.

105 Negishi, H. and Yoshihara, K. (1967) Semicontinuous channels with past history, Kŏdai Math. Sem. Rep., 19, 53-60.

106 Ovseevich, I. A. (1963) Capacities of a multipath system. Problemy Peredachi Inform. XIV, 43-58 (in Russian).

107 _____, (1968) Capacity of a randomized channel with feedback and the coordination of sources to such channels, Probl. Pered. Inform., 4, 1, 52-59. (in Russian)

106 Ovsecvich, I. A. and M. S. Pinsker (1959) The speed of transmission of information, the capacity of a multi-channel system, and its receipt using a transformation method by a linear operator, Radiotek., 3, 9-21 (in Russian).

107 _____, _____ (1961) On the capacity of a multipath system. Izv. Akad. Nauk Energet. i Avtomat. 4, 208-210 (in Russian).

110 Parthasarathy, K. R. (1963) Effective entropy rate and transmission of information through channels with additive random noise. Sankhyā, Indian J. Statist. A25, 75-84.

111 Perez, A. (1957) Sur la théorie de l'information dans le cas d'un alphabet abstrait. Trans. First Prague Conf. Information Theory, Statist. Decision Functions and Random Processes, 209-244. Prague.

112 _____ (1959) Information theory with an abstract alphabet. Generalized aspects of McMillan's theorem for the case of discrete and continuous time. Teor. Verojatnost. i Primenen. 4, 105-109 (in Russian).

113 _____(1965) Information, ε-sufficiency and data reduction problems, Kibernetika (Prague) 1, 297-323.

114 Pilk, A. (1967) Coding for source-channel pairs, MIT Quart. Progress Rep. 81, 169-173.

115 _____ (1968) Coding theorems for source-channel pairs, MIT Quart. Progress Rep. 85, 241-248.

116 Pinkston, J. T. III (1968) Block codes for discrete sources and certain fidelity criteria, MIT Quart. Progress Rep. 86, 240-246.

117 Pinsker, M. S. (1960) Information stability of Gaussian random variables and processes. Dokl. Akad. Nauk SSSR 133, 28-30 (in Russian).

115 Pinsker, M. S. (1960) The entropy, the rate of establishment of entropy and entropic stability of Gaussian random variables and and processes. Dokl. Akad. Nauk SSSR 133, 531-534 (in Russian).

119 _____ (1963) Sources of messages. Problemy Peredachi Informacii 14, 5-20 (in Russian).

120 _____ (1963) Gaussian sources. Problemy Peredachi Informacii 14, 59-100 (in Russian).

121 _____ (1965) On decoding complexity. Problemy Peredachi Inform. 1, 113-116 (in Russian) = Prob. Inform. Transm. 1, 1, (1965), 84-86.

122 _____ (1966) Some mathematical problems in the theory of information transmission, Kibernetika, (Prague), 2, 2, 117-147.

123 Powers, K., A prediction theory approach to information rates, IRE Convention Record, Vol. 4 (1956), 132-139.

124 Prelov, V. V. (1966) Asymptotic capacity of some communication channels, Probl. Pered. Inform., 2, 1, 14-27.

125 Prosser, R. T. (1966) The ϵ-entropy and ϵ-capacity of certain time-varying channels, J. Math. Anal. Appl., 16, 553-573.

126 Ratner, M. E. (1965) Asymptotic of optimal probability of error in the transmission of information over a continuous memoryless erasure symmetric channel. Problemy Kibernet. 13, 115-130 (in Russian).

127 Reiffen, B. (1962) Sequential decoding for discrete input memoryless channels. IRE Trans. Information Theory 8, 203-220.

128 _____ (1966) A per letter converse to the channel coding theorem, IEEE Trans. Inform. Theory, 12, 475-480.

129 Rényi, A. (1965) On the foundations of information theory (with discussion). Rev. Int. Statist. Inst. 33, 1-14.

130 Rényi, A. (1967) On some basic problems of statistics from the point of view of information theory. Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, 531-543

131 Rosenblatt-Roth, M. (1957) The theory of the transmission of information through statistical communication channels. Dokl. Akad. Nauk SSSR 112, 202-205 (in Russian).

132 _____ (1964) The notion of entropy in the theory of probabilities and its application in the theory of information transmission through noisy channels. Teor. Verojatnost. i Primenen. 9, 246-261 (in Russian).

133 _____ (1967) Approximations in information theory. Proc. Fifth. Berkeley Symposium on Mathematical Statistics and Probability, 1, 545-564

134 Sandberg, I. W. (1966) On the sensitivity of channel capacity for the Gaussian band limited channel, Bell. Syst. Techn. J. 45, 1475-1492.

135 Savage, J. E. (1963) Sequential decoding for an erasure channel with memory. MIT Quarterly Progress Report 69, 149-154.

136 _____ (1966) Sequential decoding - the computation problem, Bell. Syst. Techn. J., 45, 149-175.

137 _____ (1966) The distribution of the sequential decoding computation time, IEEE Trans. Inform. Theory, 12, 2, 143-147.

138 _____ (1966) A bound on the reliability of block coding with feedback, Bell. Syst. Techn. J., 45, 6, 967-977.

139 Schalkwijk, J. P. M. and Kailath, T., (1966) A coding scheme for additive noise channels with feedback Part I. No bandwidth constraint, IEEE Trans. Inform. Theory, 12, 2, 172-182.

140 Schalkwijk, J. P. M. (1966) A coding scheme for additive noise channels with feedback, Part II, Band-limited signals, IEEE Trans. Inform. Theory, 12, 2, 183-189.

141 Shannon, C. E. (1948) A mathematical theory of communication. Bell. System Tech. J. 27, 379-423, 623-656.

142 _____, "The zero capacity of a noisy channel," IRE Trans. Inf. Theory, IT-2 (1956), pp. 8-19.

143 _____, "Some geometric results in channel capacity," Nachr. Tech. Fachber., Vol. 6, (1956), pp. 13-15.

144 _____, (1957) Certain results in coding theory for noisy channels. Information and Control 1, 6-25.

145 _____ (1958) Channels with side information at the transmitter. IBM J. Res. Devel. 2, 289-293.

146 _____ (1958) A note on a partial ordering for communication channels. Information and Control 1, 390-398.

147 _____ (1959) Probability of error for optimal codes in a Gaussian channel. Bell System Tech. J. 38, 611-655.

148 _____ (1960) Coding theorems for a discrete source with a fidelity criterion. Information and Decision Processes, 93-126. R. E. Machol, Ed., McGraw-Hill Book Company, New York, N.Y.

149 _____ (1961) Two-way communication channels. Proc. Fourth Berkeley Symposium on Mathematical Statistics and Probability. 1, 611-644. University of California Press, Berkeley.

150 Shannon, C. E., Gallager, R. G. and Berlekamp, E. R. (1967) Lower bounds to error probability for coding on discrete memoryless channels I, Inform. and Control 10, 65-103; II, ibid, 10, 522-552.

151 Slepian, D., A class of binary signalling alphabets, Bell System Tech. J., 35 (1956), 203-234.

152 Shen Shih-Yi (1963) A necessary and sufficient condition for satisfaction of the information criterion in the Shannon theorem. Chinese Math. 3, 419-438 (English translation).

153 _____ (1964) The fundamental problem of stationary channels. Trans. Third Prague Conf. Information Theory, Statist. Decision Functions and Random Processes, 637-639. Prague.

154 Shen Shi-Yi (1964) The coding theorem and its converse for finite memory channels, Shuxue Yinzhan 7, 339-345 (in Chinese).

155 Stiglitz, I. G. (1963) Sequential decoding with feedback. Ph.D. Thesis, M.I.T.

156 _____ (1966) Coding for a class of unknown channels, IEEE Trans. Inform. Theory, 12, 2, 189-195; correction ibid. 13, 2, 195.

157 _____ (1967) A coding theorem for a class of unknown channels, IEEE Trans. Inform. Theory, 13, 2, 217-220.

158 _____ (1967) Iterative sequential decoding, MIT, Lincoln Lab., Preprint, MS-1917.

159 Strassen, V. (1964) Asymptotische Abschätzungen in Shannon s Informationstheorie. Trans. 3rd Prague Conf. Information Theory, Statist. Decision Functions and Random Processes 689-723. Prague.

160 _____ (1964) Meszfehler and Information, Z. Wahrscheinlichkeitsth., 2, 4, 273-305.

161 Swerling, P. (1960) Paradoxes related to the rate of transmission of information, Inform. and Control, 3, 351-359.

162 Takano, K. (1957) On the basic theorems of information theory. Ann. Inst. Statist. Math. 9, 53-77.

163 Thomasian, A. J. (1960) Error bounds for continuous channels. Proc. 4th London Symp. Information Theory, 46-60, C. Cherry, Ed. Butterworths, London.

164 Tsybakov, B. S. (1961) Shannon scheme for Gaussian message with a uniform spectrum and channel with fluctuating noise. Radiotekhnika i Electronika, 6., 649-651 (in Russian).

165 _____, "On the capacity of two-path communication channels", Radiotek. i Elektron., 4 (1959), 1117-1123. (in Russian).

166 _____, "On the capacity of channels with a large number of paths," Radiotek. i Elektron., 4 (1959), 1427-1433. (in Russian).

167 _____, "The capacity of certain multi-path channels," Radiotek. i Elektron., 4 (1959), 1602-1608. (in Russian).

168 _____ (1966) Asynchronous channels with a synchrosymbol, Probl. Pered. Inform. 2, 1, 28-36. (in Russian).

169 Vajda, I. (1966) The synchronization problem of information theory, Kybernetika (Prague), 2, 314-330.

170 Weiss, L. (1960) On the strong converse of the coding theorem for symmetric channels without memory. Quart. Appl. Math. 18, 209-214.

171 Winkelbauer, K. (1960) Communication channels with finite past history. Trans. 2nd Prague Conf. Information Theory, Statist. Decision Functions and Random Processes, 685-831. Prague.

172 _____ (1964) On discrete information sources. Trans. 3rd Prague Conf. Information Theory, Statist. Decision Functions and Random Processes, 765-830. Prague.

173 Winkelbauer, K. (1967) Axiomatic definition of channel capacity and entropy rate, Trans. Fourth Prague Conf. Infor. Theory, Statist. Decision Functions, Random Processes. Publ. House Czech. Acad. Sci., Prague, 661-705.

174 Wolfowitz, J. (1957) The coding of messages subject to chance errors. Illinois J. Math. 1, 591-606.

175 _____ (1959) Strong converse of the coding theorem for semi-continuous channels. Illinois J. Math. 3, 477-489.

176 _____ (1960) Simultaneous channels. Arch. Rational Mech. Anal. 4, 371-386.

177 _____ (1960) Strong converse of the coding theorem for the general discrete finite-memory channel. Information and Control, 3, 89-93.

178 _____ (1961) A channel with infinite memory. Proc. Fourth Berkeley Symposium on Mathematical Statistics and Probability 1, 763-767. University of California Press, Berkeley.

179 _____ (1963) On Channels without capacity. Information and Control 6, 49-54.

180 _____ (1963) The capacity of an indecomposable channel. Sankhya, Indian J. Statist. A25, 101-108.

181 _____ (1967) Approximation with a fidelity criterion. Proc. Fifth Berkeley Symp., 1, 565-573. University of California Press, Berkeley.

182 _____ (1967) Memory increases capacity. Inform. and Control, 11, 423-428.

183 Wolfowitz, J. (1968) Note on the Gaussian channel with feedback and a power constraint, Inf. and Control, 12, 71-78.

184 _____ (1968) Note on a general strong converse. Inform. and Control, 12, 1-4.

185 Wozencraft, J. M. (1957) Sequential decoding for reliable communication. IRE Convention Record, 5, 11-25.

186 Wyner, A. D. (1964) An improved error bound for Gaussian channels, Bell Syst. Techn. J., 43, 3070-375

187 _____ (1965) Capabilities of bounded discrepancy decoding, Bell Syst. Techn. J., 44, 6, 1061-1122.

188 _____ (1966) Bounds on communication with polyphase coding. Bell Syst. Techn. J., 45, 523-559.

189 _____ (1966) Capacity of the product of channels, Inform. and Control, 9, 423-430.

190 _____ (1967) On the probability of error for communication in white Gaussian noise, IEEE Trans. Inf. Theory, 13, 86-90.

191 _____ (1967) Random packings and coverings of the unit n-sphere, Bell. Syst. Techn. J., 46,

192 _____ (1968) On the Schalkwijk-Kailath coding scheme with a peak energy constraint, IEEE Trans. Inf. Theory, 14, 1, 129-134.

193 _____ (1968) Communication of analog data from a Gaussian source over a noisy channel, Bell. Syst. Techn. J., 47 , 801-812.

194 Yoshihara, Ken-Ichi (1964) Simple proofs for the strong converse theorems in some channels. Kodai Math. Sem. Rep. 16, 213-222.

195 _____ (1965) Coding theorems for the compound semi-continuous memoryless channels, Kodai Math. Sem. Rep. 17, 1, 30-43.

196. Yoshihara, Ken-Ichi (1966) The Shannon problem for a unique transmission method, Yokahama University Sci. Rep. Sect. 1, 13, 1-28.

197. Yudkin, H.L. (1967) On the exponential error bound and capacity for finite state channels, MIT, Lincoln Lab., JA-3036, DS-4708, presented at the Intern. Symp. on Inform. Theory, Sept. 1967, San Remo, Italy.

198. Zhang, Zhao-Zhi. (1965) Some results obtained with almost-periodic channels, Acta Math. Sinica, 15, 127-135 (Chinese); transl. as Chinese Math. - Acta 6 (1965), 428-436.

199. Zigangirov, K. Sh. (1968) Sequential decoding with the random-coding exponent of the probability or error, Probl. Pered. Inform., 4, 2, 83-85.

200. Ziv, J. (1963) Successive decoding scheme for memoryless channels. IEEE Trans. Information Theory 9, 97-104.

201. _____(1965) Probability of decoding error for random phase and Rayleigh fading channels. IEEE Trans. Inform. Theory, 11, 53-6.

202. _____(1966) Further results on the asymptotic complexity of an iterative coding scheme, IEEE Trans. Inform. Theory, 12, 168-171.

203. _____(1967) Asymptotic performance of complexity of a coding scheme for memoryless channels, IEEE Trans. Inform. Theory, 13, 3, 356-359.

END